

Évaluation : vade-mecum et impact clinique

Jean-Michel CHABOT
Faculté de médecine - Marseille

L'évaluation est dans le secteur de la santé, tout comme dans celui de l'éducation et de la formation, une préoccupation croissante.

Ainsi, il est aujourd'hui acquis que toute action inclut sa propre évaluation, ou plus simplement que l'évaluation d'une action fait partie intégrante de cette action, ou encore plus radicalement : il n'y a pas d'action sans évaluation. Dès lors, il devient licite de s'intéresser à l'évaluation, à ses qualités, à ses caractéristiques et à ses finalités ; toutes choses qui ont été trop longtemps méconnues ou négligées. D'une manière quasi unanime,¹⁻⁴ trois notions principales sont avancées, dès lors que l'on discute des qualités que doit présenter, autant que possible, toute évaluation.

Ces trois qualités sont sa validité, sa commodité et sa robustesse. Derrière ces trois termes, qui empruntent volontiers au langage familier, se cachent une série de caractéristiques métrologiques, le plus souvent d'inspiration anglo-saxonne. Traduites en français, ces caractéristiques métrologiques s'expriment par un ensemble de néologismes, inégalement compréhensibles et quelquefois irritants pour le lecteur non spécialiste.

Parmi ces trois qualités, la commodité s'entend aisément. Elle renvoie à la faisabilité (par l'évaluateur) et à l'acceptabilité (par l'évalué) que doit rechercher toute évaluation. Force est de reconnaître que cette commodité est trop souvent négligée.

La robustesse est également facilement décodable : elle regroupe sous ce seul terme la fiabilité, l'objectivité, la reproductibilité que chacun est en droit d'espérer de toute évaluation, d'autant que des décisions seront susceptibles d'en découler.

La validité reste la notion la plus délicate à préciser ; c'est pourtant la qualité première d'une évaluation. Cette délicatesse tient sans doute au caractère polysémique que la langue française confère à ce substantif et à l'adjectif correspondant. On peut également préciser que validité et pertinence sont deux termes qui recouvrent sensiblement la même notion. Et ainsi convenir que la validité est pour un programme d'évaluation ce que la pertinence signifie pour la formation.

En pratique, en matière d'évaluation, la validité apprécie dans quelle mesure une politique, une procédure, un instrument ou un critère (d'évaluation) permettent effectivement « de mesurer ce qu'ils ont pour objet de mesurer ».

La validité d'une évaluation rend donc compte de sa capacité à bien traduire ce qu'il importe d'analyser. Dans cette mesure, il est préférable de renoncer à une évaluation dont la validité serait trop faible ou même incertaine, d'autant que sa commodité serait elle-même médiocre.

Au-delà des trois notions de validité, commodité et robustesse, il reste quatre caractéristiques à exposer pour être tout à fait à l'aise avec l'évaluation : ces caractéristiques sont portées par les termes d'évaluation « formative » ou « sommative » et d'évaluation « critériée » ou « normée » (ainsi que différentes combinaisons possibles entre ces quatre termes).

Selon le modèle le plus communément admis, l'évaluation comporte trois temps successifs : d'abord le recueil des

données ; ensuite la formulation de l'appréciation (jugement) ; et, le cas échéant, la prise de décision résultante. Sous ce même modèle, il faut distinguer deux types d'évaluation, principalement distinctes en raison de l'utilisation qui en est faite (c'est-à-dire leur finalité). Ces deux types sont l'évaluation formative et l'évaluation sommative.

L'évaluation formative est plutôt étrangère à notre culture ; c'est ce qui fait sa difficulté. Elle consiste – comme toute évaluation – à obtenir des données (qui traduisent l'objet à évaluer) de manière à porter une appréciation. Cependant, dans l'évaluation formative, cette appréciation n'a pour seule finalité que de favoriser un progrès permanent, à l'exclusion de toute autre utilisation, notamment sommative ou sanctionnante. Ainsi, l'évaluation formative (*formative assessment* des Anglo-Saxons) se traduit comme une approche dynamique, axée vers l'amélioration continue de l'objet évalué.

De son côté, l'évaluation sommative – dont la synonymie riche reconnaît les termes de « finale » ou « certifiante » ou « sanctionnante » – a pour finalité la prise de décision au terme d'une procédure d'évaluation. Ainsi, l'évaluation sommative se traduit par le constat d'un état à un moment donné (vision statique ou stabilisée). C'est ce type d'évaluation – et les menaces qui s'y rattachent en cas de résultat insuffisant – qui se trouve totalement associé à notre culture dominante ; d'où la difficulté d'envisager sans crainte l'évaluation.

Les deux dernières caractéristiques d'une évaluation concernent son cadre « critérié » ou au contraire « normé ». Identifier un (ensemble de) critère(s), puis recueillir des données, avant de les apprécier par rapport à ces critères, voilà qui définit l'évaluation critériée. On peut ainsi, par exemple, identifier l'ensemble des critères qui constituent une prévention du risque cardiovasculaire, puis analyser (apprécier) les données d'activité d'un ou plusieurs médecins pendant un temps d'observation plus ou moins long. Cette évaluation critériée permet de situer le comportement professionnel de chacun ou du groupe des médecins observés, par rapport à la liste des critères prédéterminés. D'une toute autre essence est l'évaluation normée, où le but n'est pas de situer les objets évalués par rapport à des critères prédéterminés, mais les uns par rapport aux autres ; le but est donc d'établir un classement (une hiéar-

chie), sans se soucier de la valeur intrinsèque de chacun des objets évalués.

Le propos n'est pas d'opposer évaluation normée et évaluation critériée et d'en privilégier une par rapport à l'autre ; le propos est simplement de les distinguer, en observant qu'elles n'ont pas la même finalité et en constatant que l'on passe souvent – à son insu – de l'une à l'autre, ce qui peut provoquer des confusions.

De nombreux exemples traduisant la mise en œuvre de ces évaluations dans le secteur de la santé peuvent être donnés. En particulier les deux suivants, porteurs d'impacts considérables sur l'exercice clinique quotidien des médecins et l'état des malades. Évidemment, il s'agit d'évaluations de forte validité et de très bonne commodité...

La première⁵ concerne la réalisation des coloscopies de dépistage et le taux d'adénomes identifiés qui en résulte. Trois critères simples exprimant la qualité de la coloscopie (temps de retrait du coloscope, préparation du côlon et amplitude de la visualisation du cæcum au rectum) et donc l'utilité clinique qui en découle ont ainsi été identifiés/confirmés.

La seconde⁶ concerne l'indication des β -bloquants en post-infarctus, dont l'utilité avait été démontrée il y a 25 ans (*evidence based...*) mais dont la prescription effective restait insuffisante (encore moins de 60 % des malades éligibles en 1998).

Les autorités médicales américaines viennent d'indiquer que la quasi-totalité des malades éligibles recevaient un β -bloquant – enfin – et qu'en conséquence l'analyse de cet indicateur allait être suspendue, l'objectif étant atteint. De telles évaluations peuvent parfaitement être conduites dans un cadre formatif (amélioration de la qualité) ou sommatif (p. ex. politique d'allocation de ressources). C'est aux professionnels et aux tutelles responsables d'en décider. ■

1. Dauphinee WD. Assessing clinical performance. Where do we stand and what might we expect? JAMA 1995;274:741-3.
2. Landon BE, Normand SL, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. JAMA 2003;290:1183-9.
3. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med 2005;142:260-73.
4. Campbell SM, Braspenning J, Hutchinson A, Marshall MN. Research methods used in developing and applying quality indicators in primary care. BMJ 2003;326:816-9.
5. Lieberman D. Measuring the quality of colonoscopy. N Engl J Med 2006;355:2188-9.
6. Lee TH. Eulogy for a quality measure. N Engl J Med 2007;357:1175-7.